Game of Missuggestions
Semantic Analysis of Search-Autocomplete Manipulations

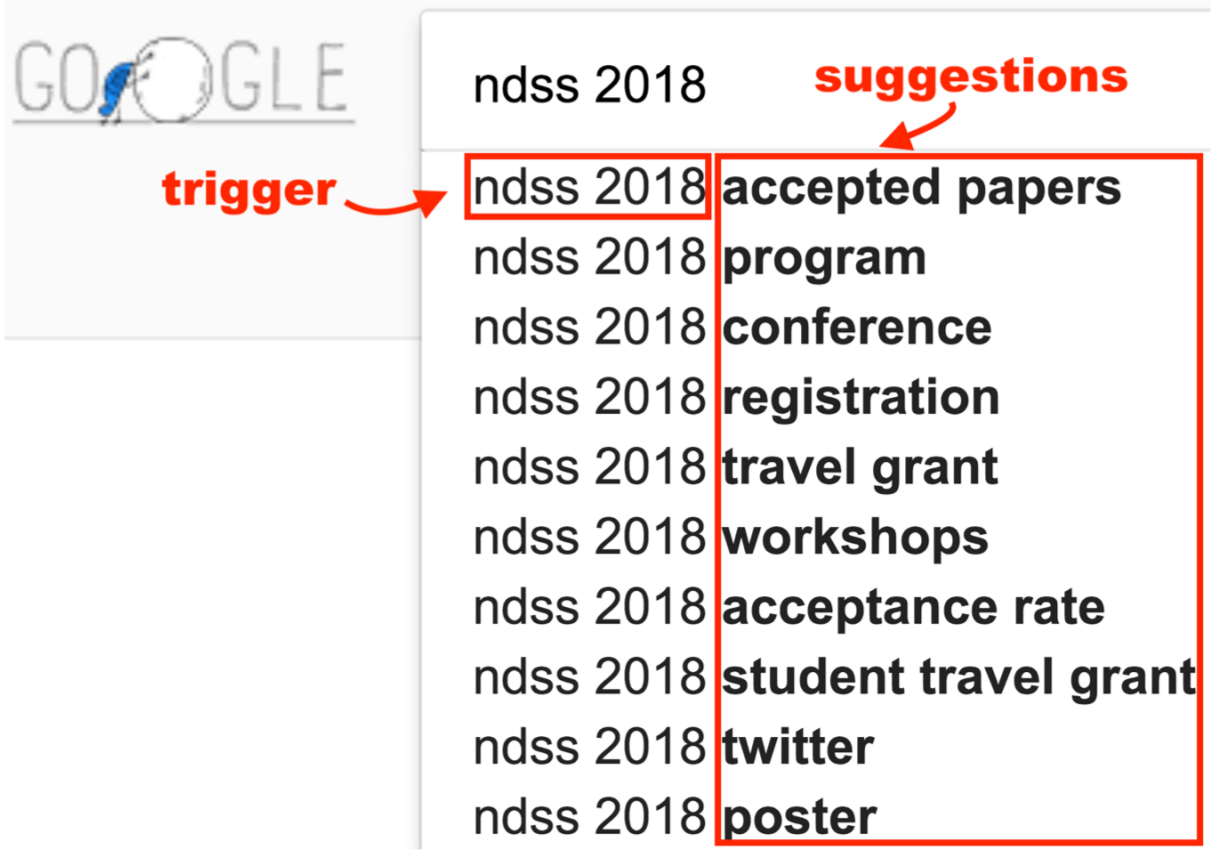
Peng Wang[1], Xianghang Mi[1], Xiaojing Liao[2],
XiaoFeng Wang[1], Kan Yuan[1], Feng Qian[1], Raheem Beyah[3]

Indiana University Bloomington[1]

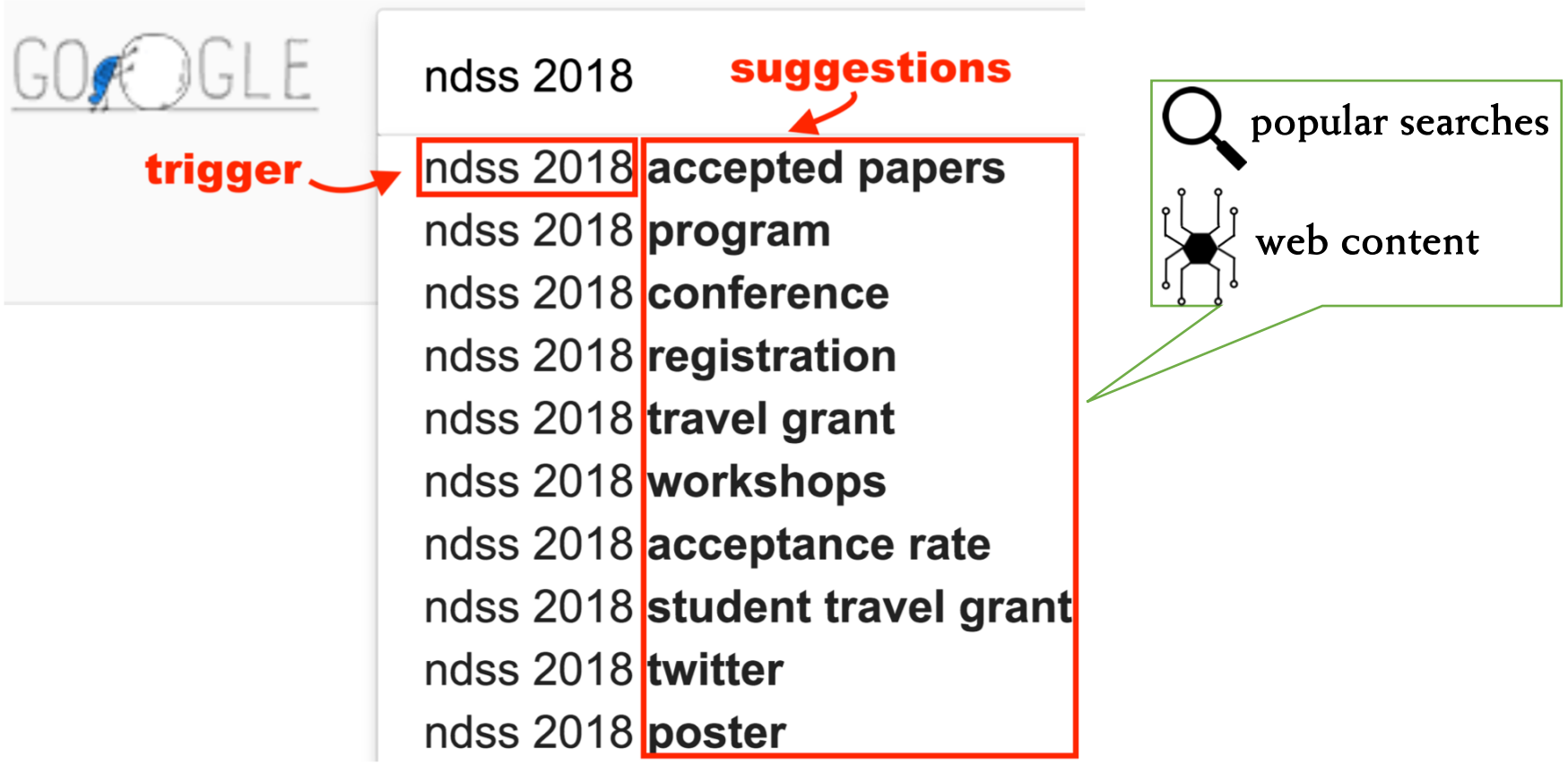
William and Mary[2]

Georgia Institute of Technology[3]

Autocomplete

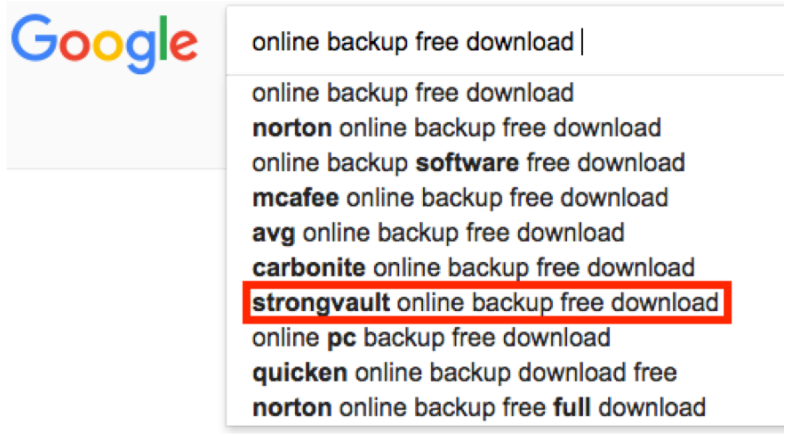


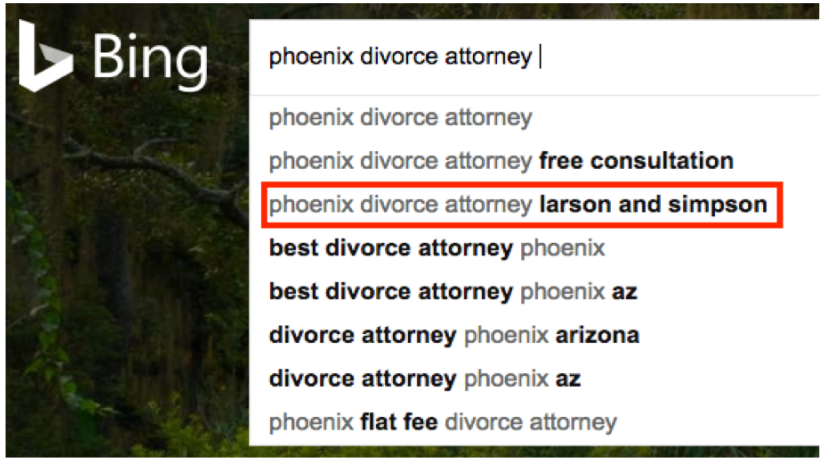
Autocomplete

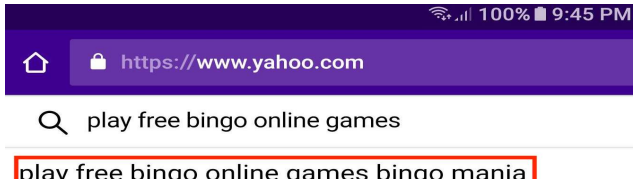


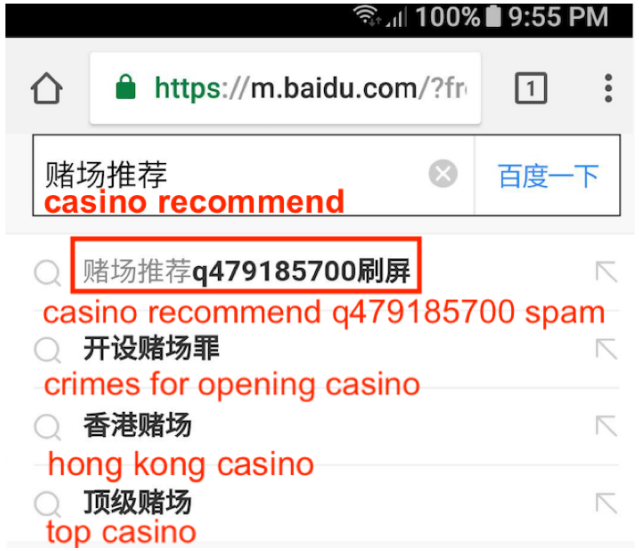
How predictions are made
https://support.google.com/websearch/answer/106230

Winter is here

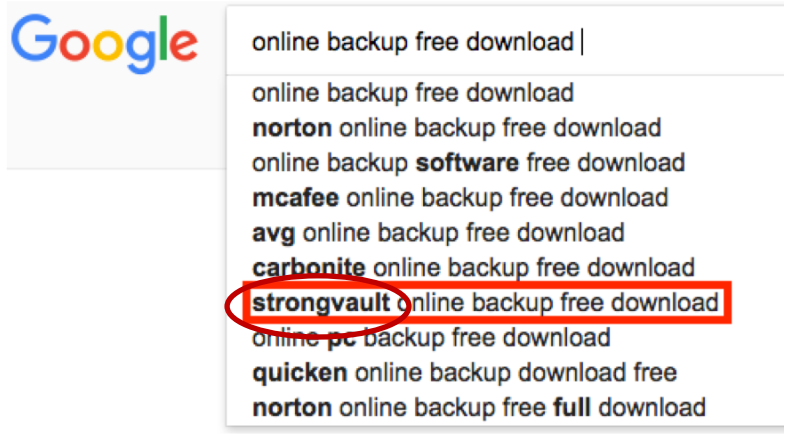




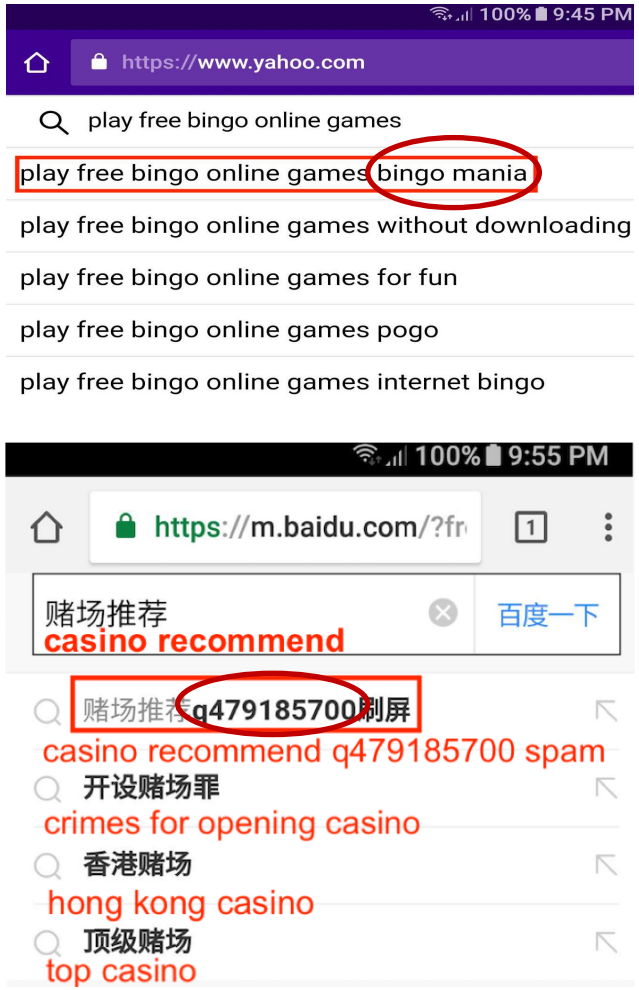


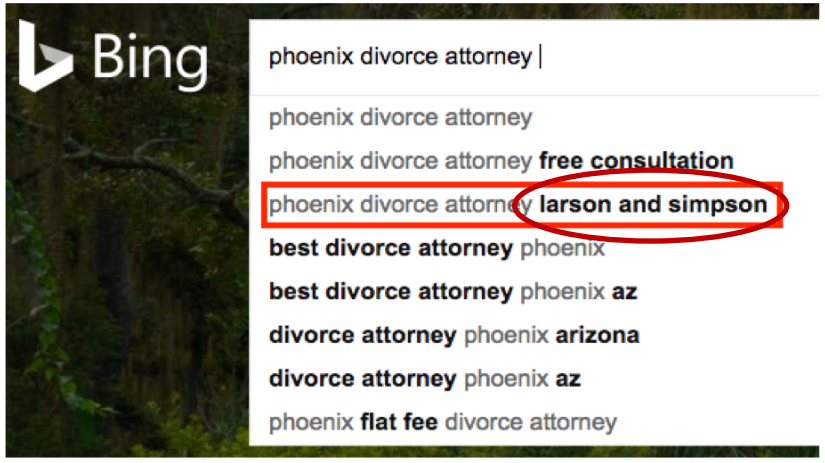


Winter is here



promotion target





GAME OF MISSUGGESTIONS

Autocomplete Manipulation

pollute search logs

Service Provider	Country	Supported Platforms	Cost	Time
Affordable Reputation Management	USA	Google, Bing, Yahoo!, Amazon etc.	$370 / m	1 - 3 months
Search Reputation	USA	Google, Bing, Yahoo!	$1,200 / m	3 months
Google Autocomplete Changer	USA	Google, Bing, Yahoo!	$2,500 (fixed)	3 months
Seopmb	China	Baidu, Sogou, 360	$1 ~$20 / d	1 month
iXiala	China	Baidu, Sogou, 360, Taobao etc.	$2 ~$12 / d	3 - 15 days

Autocomplete Manipulation

pollute search logs

Service Provider	Country	Supported Platforms	Cost	Time
Affordable Reputation Management	USA	Google, Bing, Yahoo!, Amazon etc.	$370 / m	1 - 3 months
Search Reputation	USA	Google, Bing, Yahoo!	$1,200 / m	3 months
Google Autocomplete Changer	USA	Google, Bing, Yahoo!	$2,500 (fixed)	3 months
Seopmb	China	Baidu, Sogou, 360	$1 ~$20 / d	1 month
iXiala	China	Baidu, Sogou, 360, Taobao etc.	$2 ~$12 / d	3 - 15 days

GAME OF MISSUGGESTIONS

Autocomplete Manipulation

pollute search logs

Service Provider	Country	Supported Platforms	Cost	Time
Affordable Reputation Management	USA	Google, Bing, Yahoo!, Amazon etc.	$370 / m	1 - 3 months
Search Reputation	USA	Google, Bing, Yahoo!	$1,200 / m	3 months
Google Autocomplete Changer	USA	Google, Bing, Yahoo!	$2,500 (fixed)	3 months
Seopmb	China	Baidu, Sogou, 360	$1 ~$20 / d	1 month
iXiala	China	Baidu, Sogou, 360, Taobao etc.	$2 ~$12 / d	3 - 15 days

pollute web content

compromised websites

spam hosting webpages

Autocomplete Manipulation

pollute search logs

Service Provider	Country	Supported Platforms	Cost	Time
Affordable Reputation Management	USA	Google, Bing, Yahoo!, Amazon etc.	$370 / m	1 - 3 months
Search Reputation	USA	Google, Bing, Yahoo!	$1,200 / m	3 months
Google Autocomplete Changer	USA	Google, Bing, Yahoo!	$2,500 (fixed)	3 months
Seopmb	China	Baidu, Sogou, 360	$1 ~$20 / d	1 month
iXiala	China	Baidu, Sogou, 360, Taobao etc.	$2 ~$12 / d	3 - 15 days

pollute web content

compromised websites

spam hosting webpages

Challenges

- Search log analysis
  can only be done by search providers

- Web content analysis
  a thorough study is non-trivial on massive data

Little understanding about the real-world impacts of illicit promotions

Sacabuche

Search AutoComplete Abuse Checking

- **first** detection system **without** accessing to search logs
- **novel NLP techniques** achieves highly efficient, accurate and scalable
- **first large-scale analysis** of autocomplete missuggestions
- **first step** to understand the ecosystem of this underground business

Observation

- Semantic inconsistency

  trigger: online backup free download

  legitimate: ✓ online backup **software** free download

  manipulated: ✗ **strongvault** online backup free download

Observation

- Semantic inconsistency

trigger: online backup free download

semSim=0.96

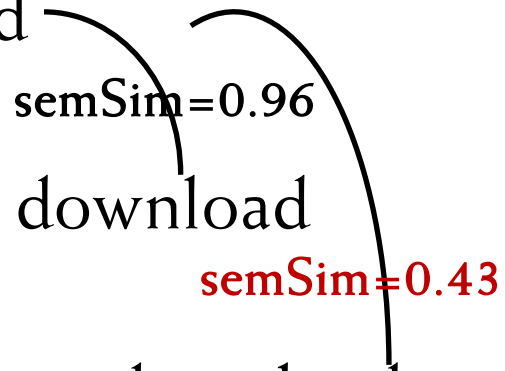
legitimate: ✓ online backup **software** free download

manipulated: ✗ **strongvault** online backup free download

Observation

- Semantic inconsistency

trigger: online backup free download



legitimate: ✓ online backup **software** free download

manipulated: ✗ **strongvault** online backup free download

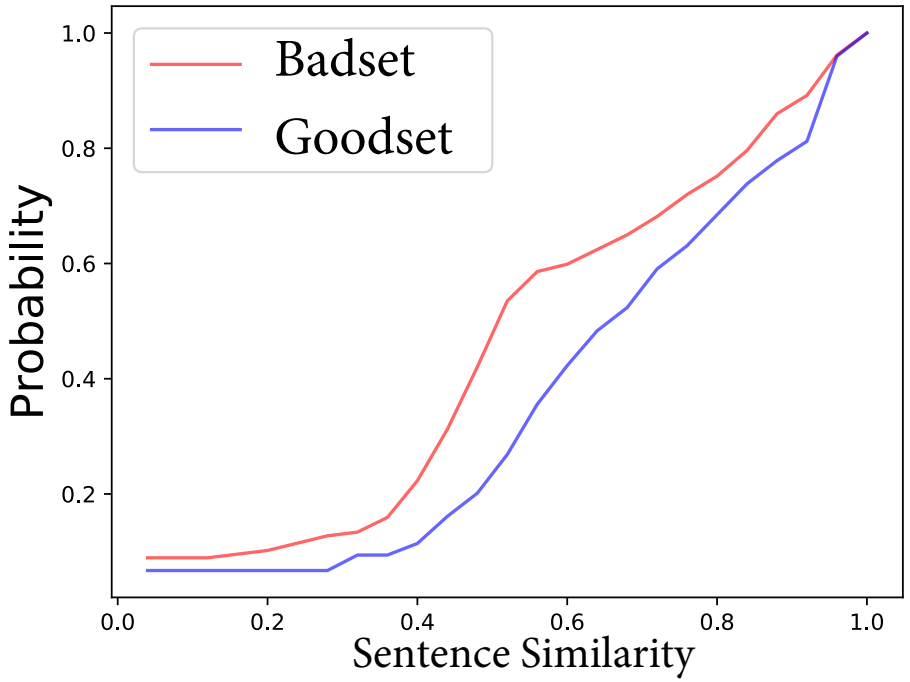
Sentence Similarity

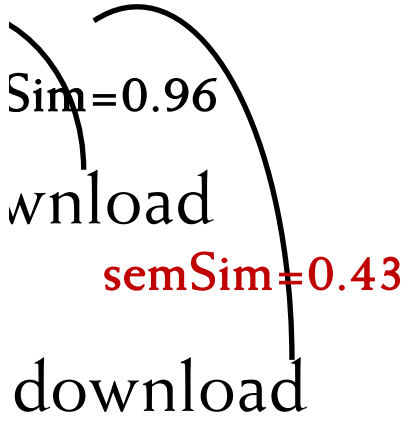
- Semantic inconsistency

  trigger:

  legitimate:

  manipulated:

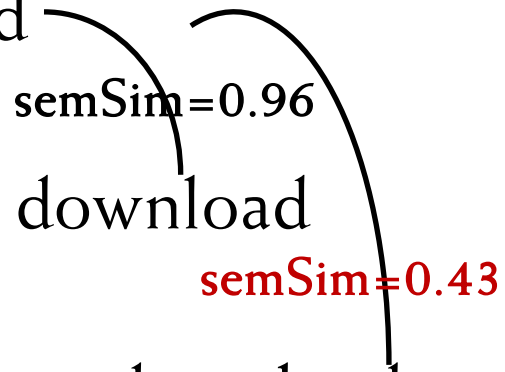




Observation

- Semantic inconsistency

trigger: online backup free download



legitimate: ✓ online backup **software** free download

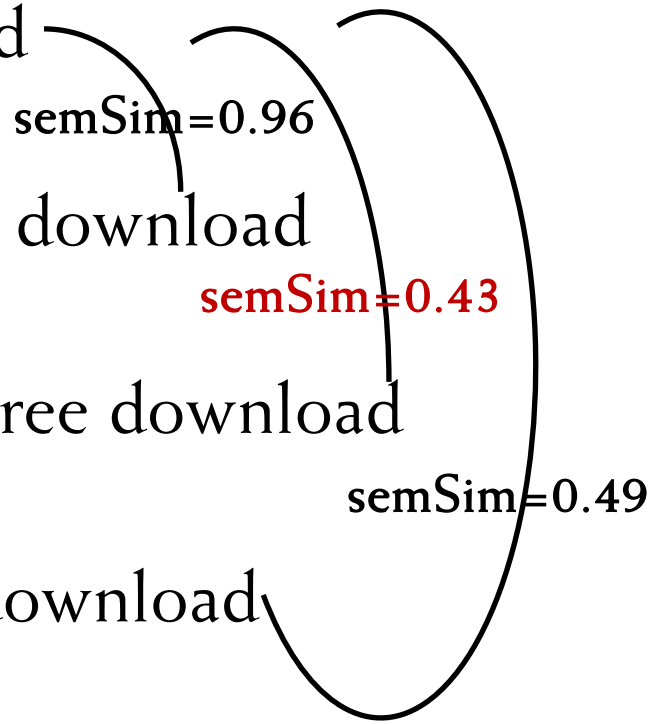
manipulated: ✗ **strongvault** online backup free download

legitimate: ✓ **norton** online backup free download

Observation

- Semantic inconsistency

trigger: online backup free download



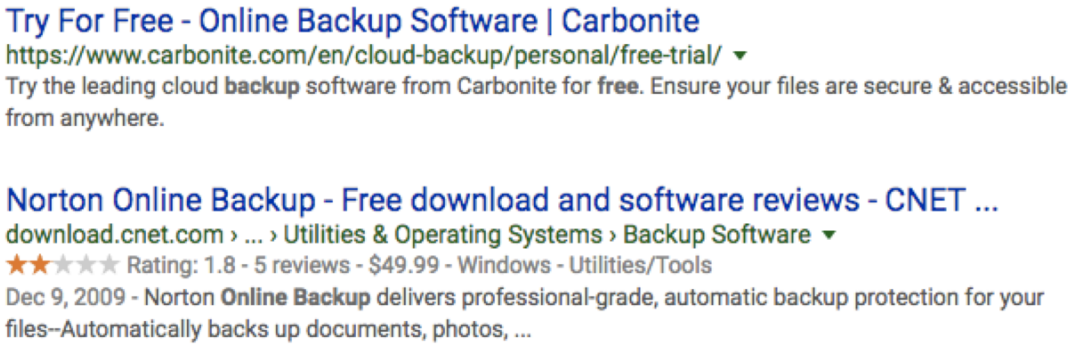
legitimate: ✓ online backup **software** free download

manipulated: ✗ **strongvault** online backup free download

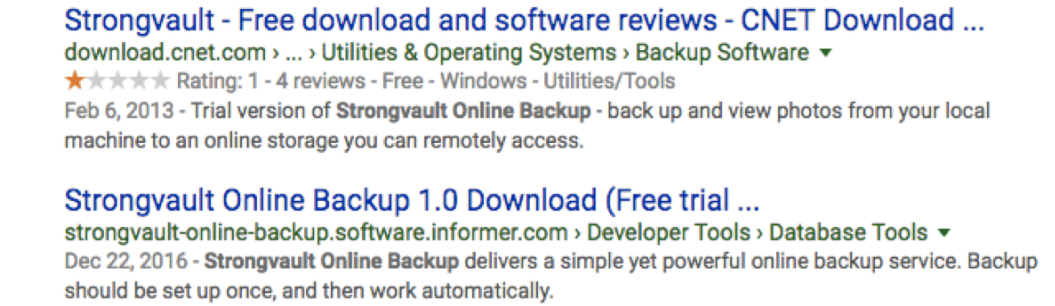
legitimate: ✓ **norton** online backup free download

Observation

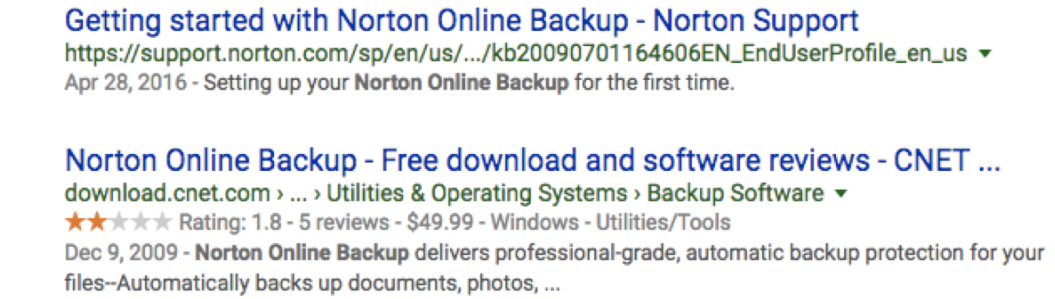
- Search results inconsistency



trigger: online backup free download



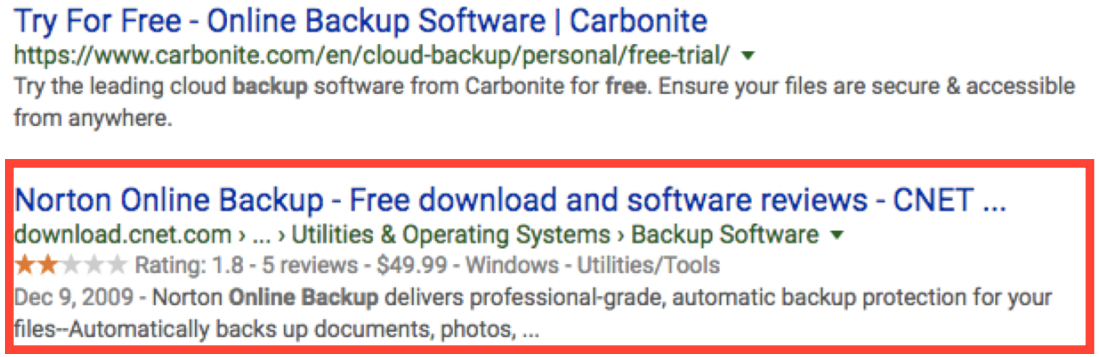
missuggestion: **stongvault** online backup free download



suggestion: **norton** online backup free download

Observation

- Search results inconsistency



trigger: online backup free download

missuggestion: **stongvault** online backup free download

suggestion: **norton** online backup free download

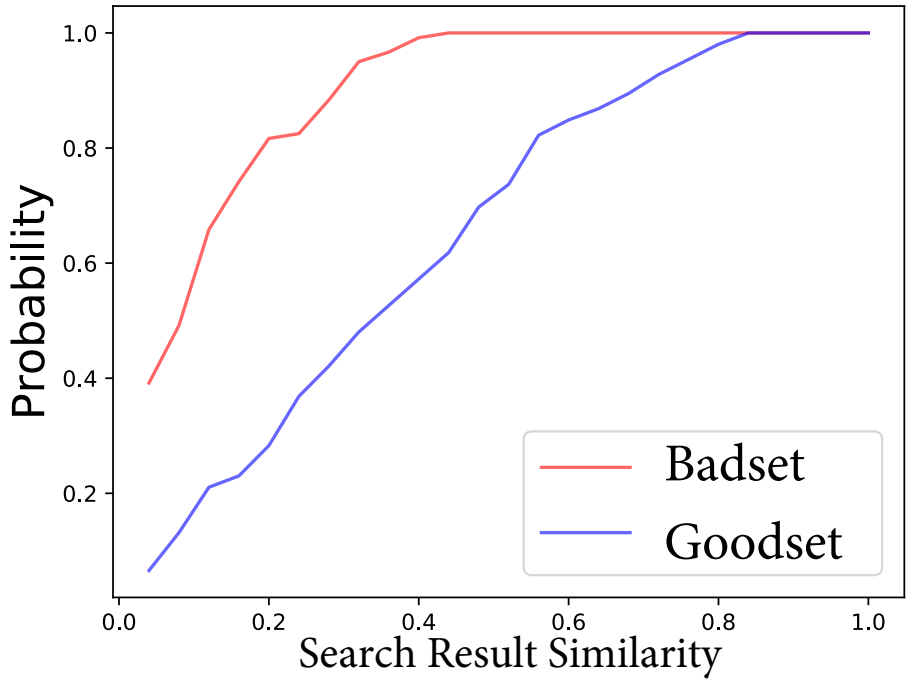
Search Results Similarity

- Search results inconsistency

Try For Free - Online Backup Software | C
https://www.carbonite.com/en/cloud-backup/perso
Try the leading cloud **backup** software from Carbonite fo
from anywhere.

Norton Online Backup - Free download a
download.cnet.com › ... › Utilities & Operating Syster
★★☆☆☆ Rating: 1.8 - 5 reviews - $49.99 - Windows - U
Dec 9, 2009 - Norton **Online Backup** delivers professiona
files--Automatically backs up documents, photos, ...

trigger: online backup



Strongvault - Free download and software reviews - CNET Download ...
download.cnet.com › ... › Utilities & Operating Systems › Backup Software
★☆☆☆☆ Rating: 1 - 4 reviews - Free - Windows - Utilities/Tools
Online Backup - back up and view photos from your local
notely access.

) Download (Free trial ...
former.com › Developer Tools › Database Tools
delivers a simple yet powerful online backup service. Backup
omatically.

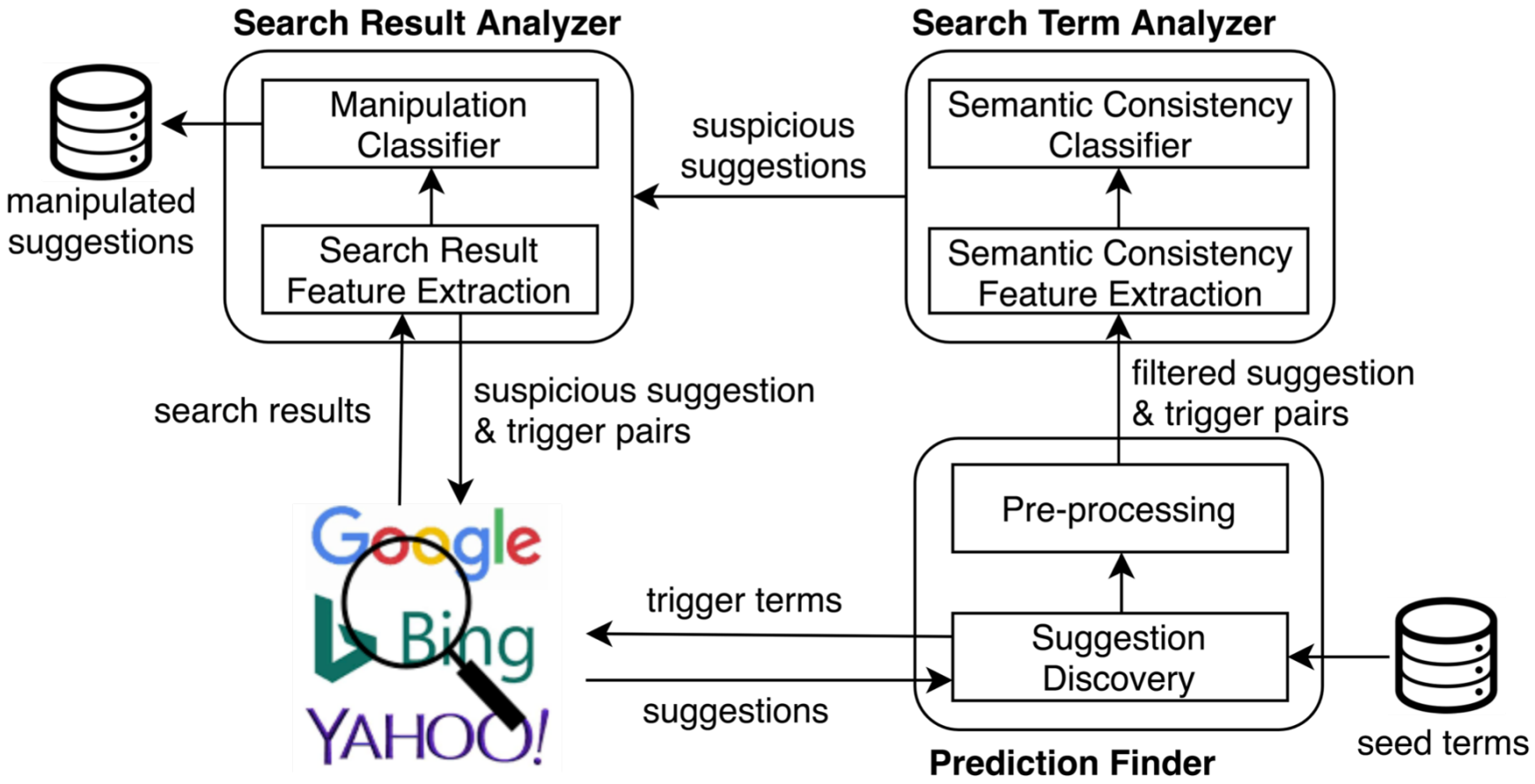
ault online backup free download

nline Backup - Norton Support
'.../kb20090701164606EN_EndUserProfile_en_us
ine Backup for the first time.

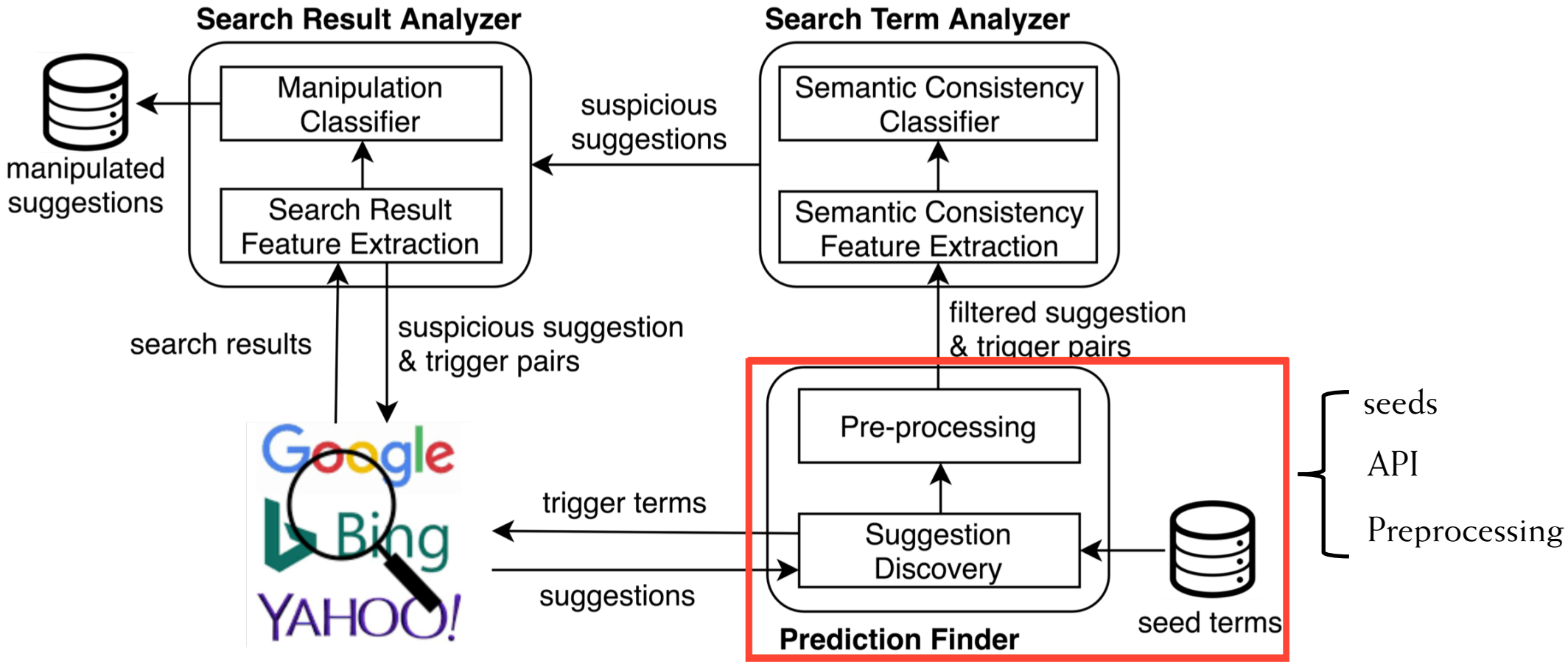
lownload and software reviews - CNET ...
erating Systems › Backup Software
9 - Windows - Utilities/Tools
ers professional-grade, automatic backup protection for your
, photos, ...

nline backup free download

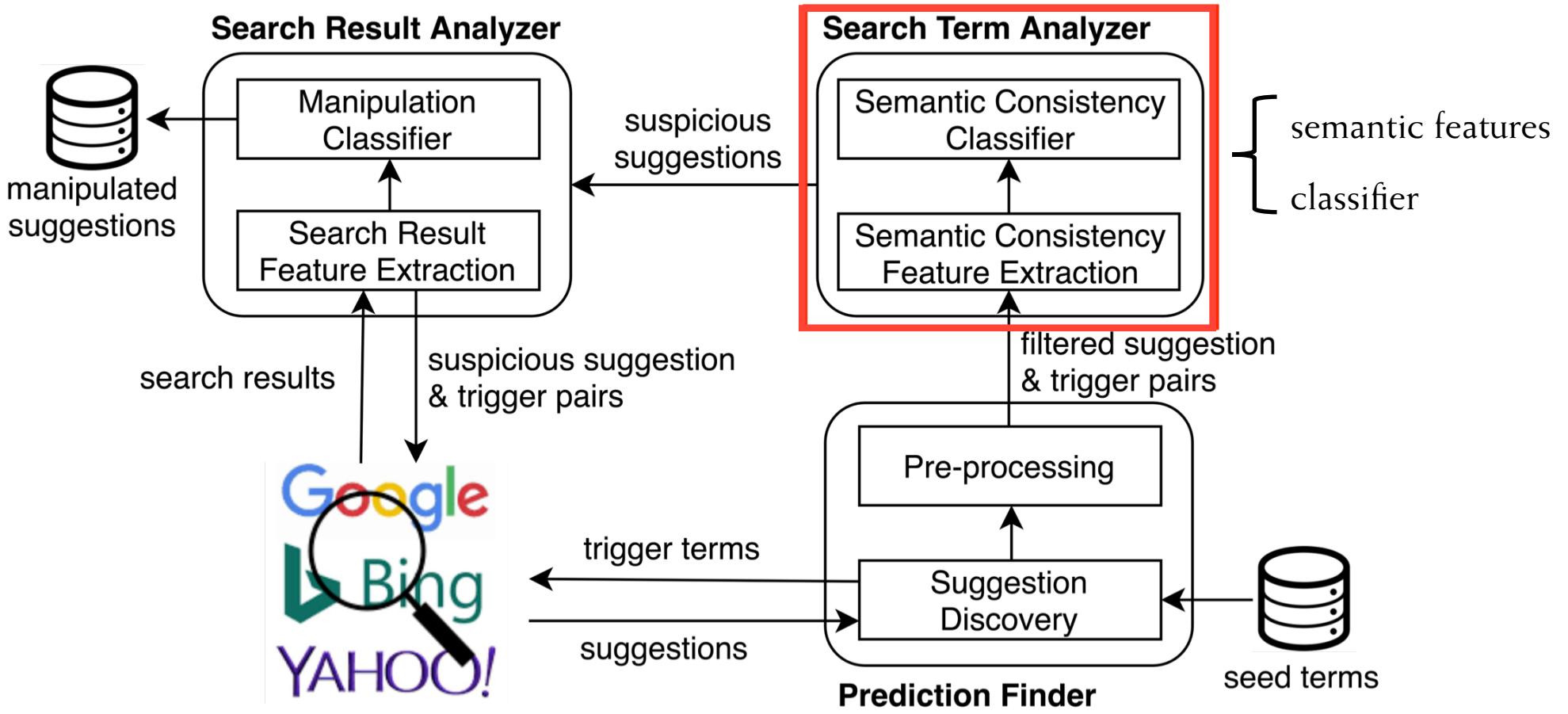
Architecture



Prediction Finder



Search Term Analyzer



Semantic Feature example

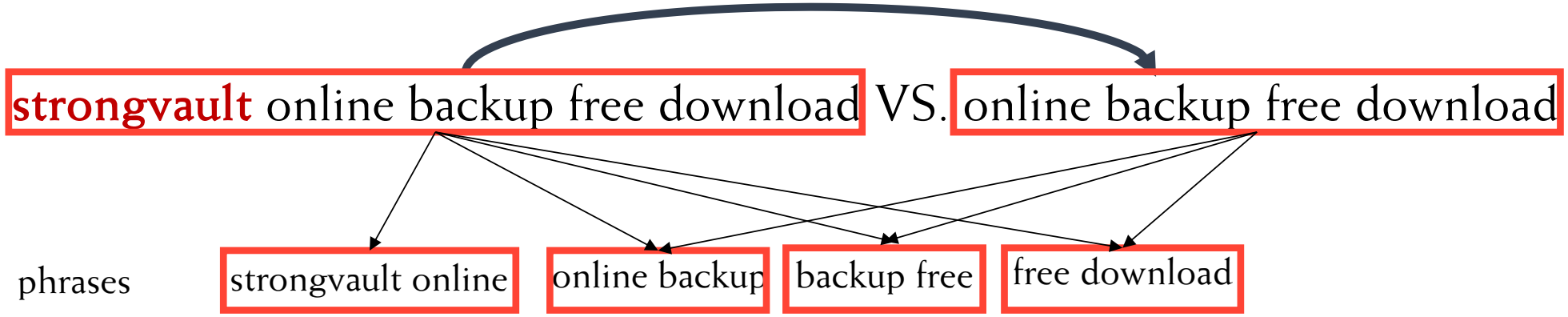
online backup free download -> **strongvault** online backup free download

- Sentence level similarity

Semantic Feature example

online backup free download -> **strongvault** online backup free download

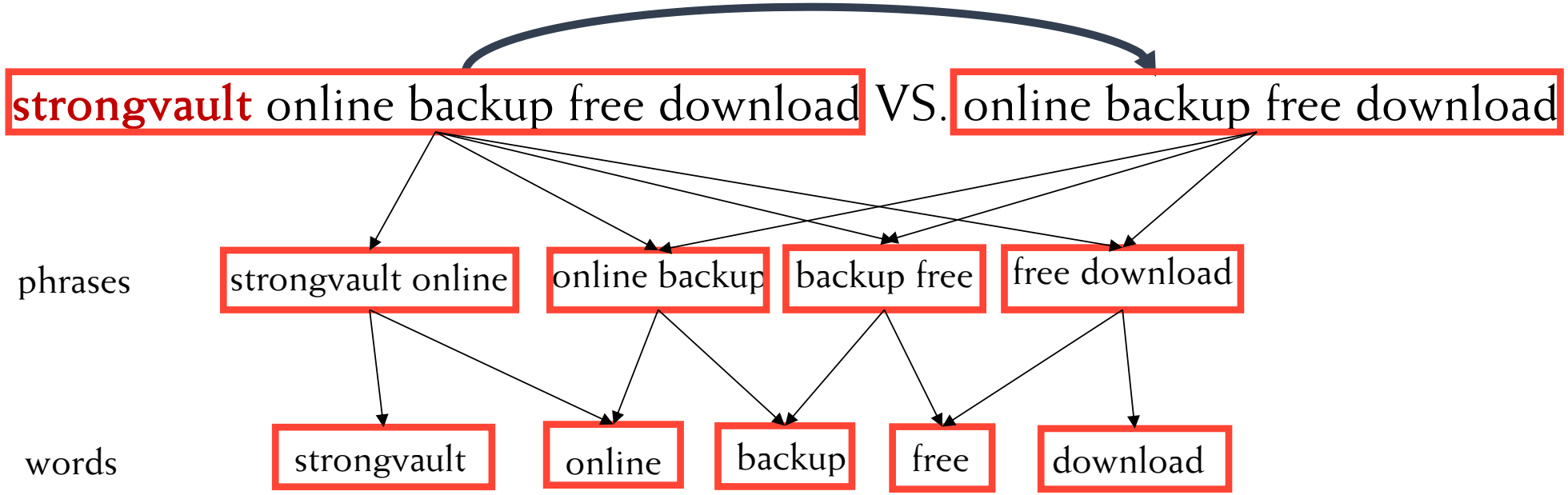
- Sentence level similarity



Semantic Feature example

online backup free download -> **strongvault** online backup free download

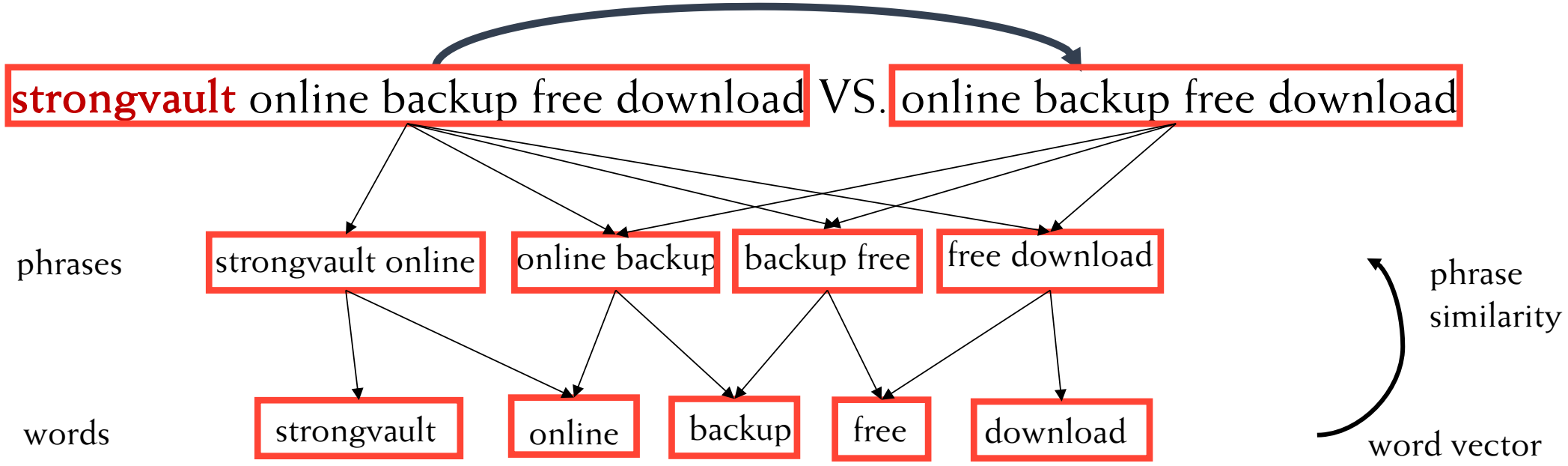
- Sentence level similarity



Semantic Feature example

online backup free download -> **strongvault** online backup free download

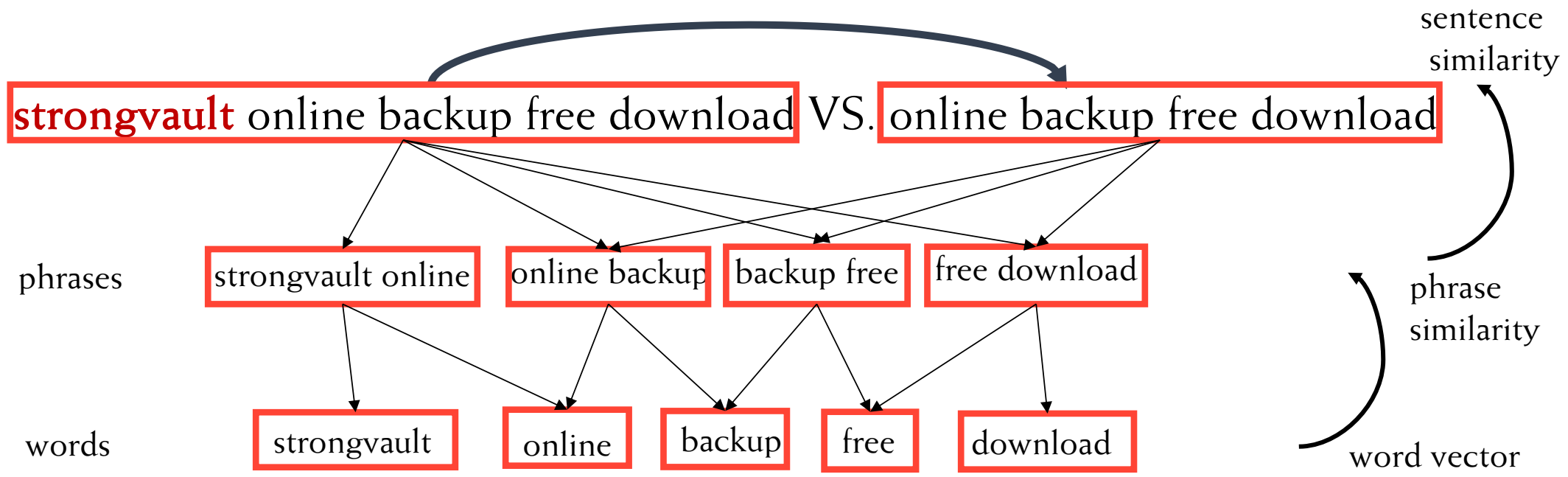
- Sentence level similarity



Semantic Feature example

online backup free download -> **strongvault** online backup free download

- Sentence level similarity



Semantic Features

- Sentence similarity

$$F_{ss}(s^a, s^t) = \frac{SK(s^a, s^t)}{\sqrt{SK(s^a, s^a)SK(s^t, s^t)}}, \qquad SK(s^a, s^t) = \sum \lambda^2 PK(p^a, p^t)$$

$$PK(p^a, p^t) = \prod_{i=1}^{len} WK\left(w_i^a, w_i^t\right), \qquad WK\left(w^i, w^j\right) = \left[\frac{1}{2}(1 + \cos Sim(w_i, w_j))\right]^{\alpha}$$

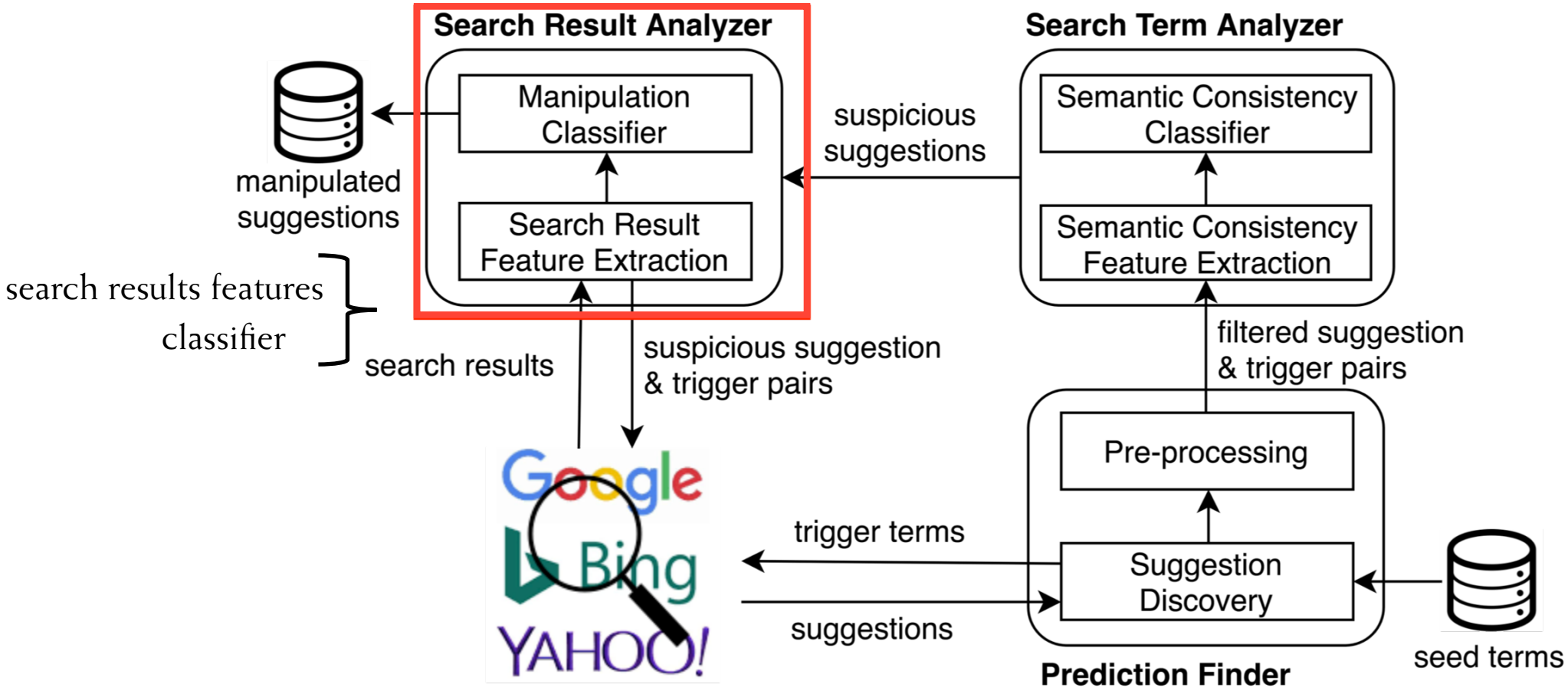
- Word similarity

$$F_{ws}(w^a, w^t) = MAX(AVG_j\left(WK\left(w_i^a, W_j^i\right)\right))$$

- Infrequency

$$F_{if}(w^a, w^t) = \frac{MAX_j(9 - log_{10}Freq\left(w_j^t\right))}{MAX_i(9 - log_{10}Freq\left(w_i^a\right))}$$

Search Result Analyzer



Search Result Features

- Result similarity

  $F_{rs}(D^a, D^t) = (1-p)\sum_{d=1}^{\infty} p^{d-1} A\,(D^t, D^a)^d$

- Content impact

  $F_{ci}(w^a, H^a, H^t) = \ MIN_i(R(w^a, H^a, H^t))$

- Result popularity

  $F_{\mathrm{rp}}(D^a, D^t) = RBO(AP^a(D^a), AP^t(D^t))$

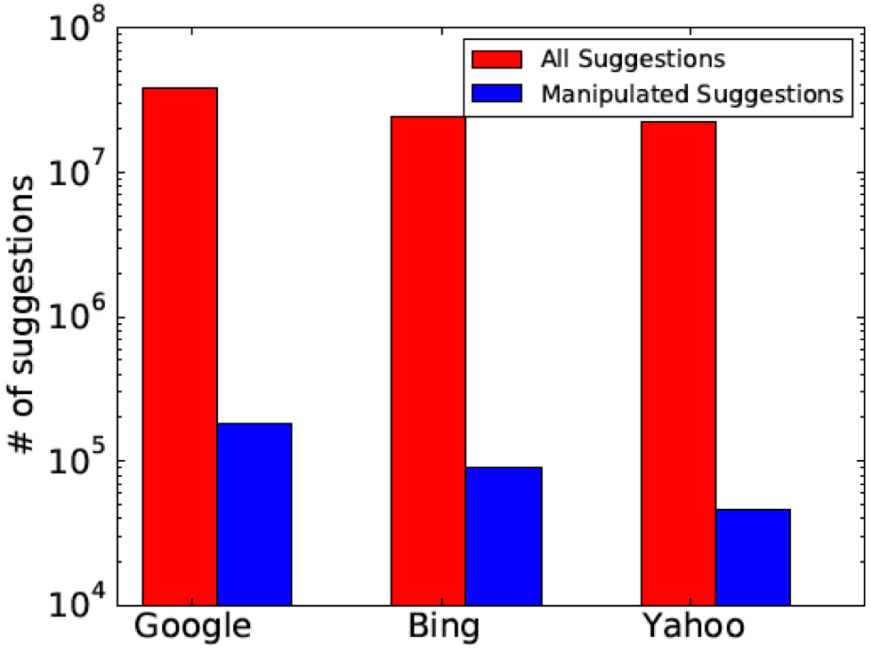
- Result size

  $F_{rs}(N^a, N^t) = \dfrac{N^a - N^t}{N^t}$

Evaluation

- Datasets
  - Badset: 150 missuggestions, 296 result pages
  - Goodset: 300 legitimate suggestions, 593 result pages
  - Unknown set: 114 millions trigger-suggestion pairs, 1.6 millions result pages
- Accuracy and coverage
  - Ground truth: precision 96.23%, recall 95.63%
  - Unknown set: precision 95.4% on 1K suspicious trigger-suggestion pairs
- Performance
  - 1.5s / trigger-suggestion pair

Scope and magnitude

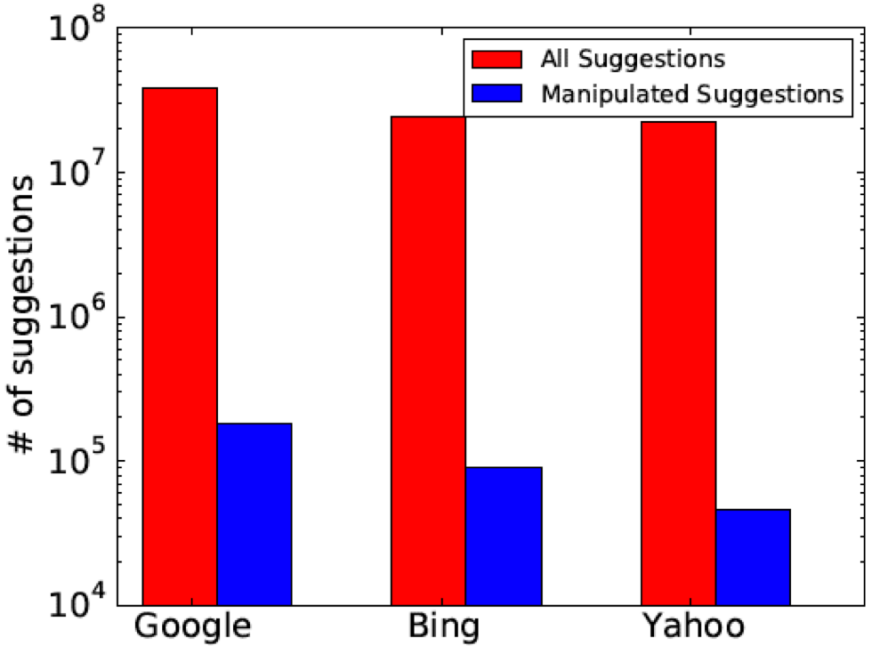


Number of missuggestions on each platform
(G: 0.48%, B: 0.37%, Y: 0.2%)

Trigger Category	# of terms	# of manipulations	Vol-ume
Lending Products	1389	1580/34629	4.13%
Home Services	17059	16712/413836	2.47%
Pharmaceutical	3715	3876/93929	2.09%
Auto Services	3477	3916/92220	1.08%
Technology	29115	32696/762548	0.91%
Education	17311	17628/423352	0.81%
Shopping	30465	32986/773087	0.72%
Gambling	413	454/10314	0.68%
Travel	6434	6554/153842	0.48%
Legal Services	14064	14084/348548	0.55%

Categories of the polluted triggers

Scope and magnitude



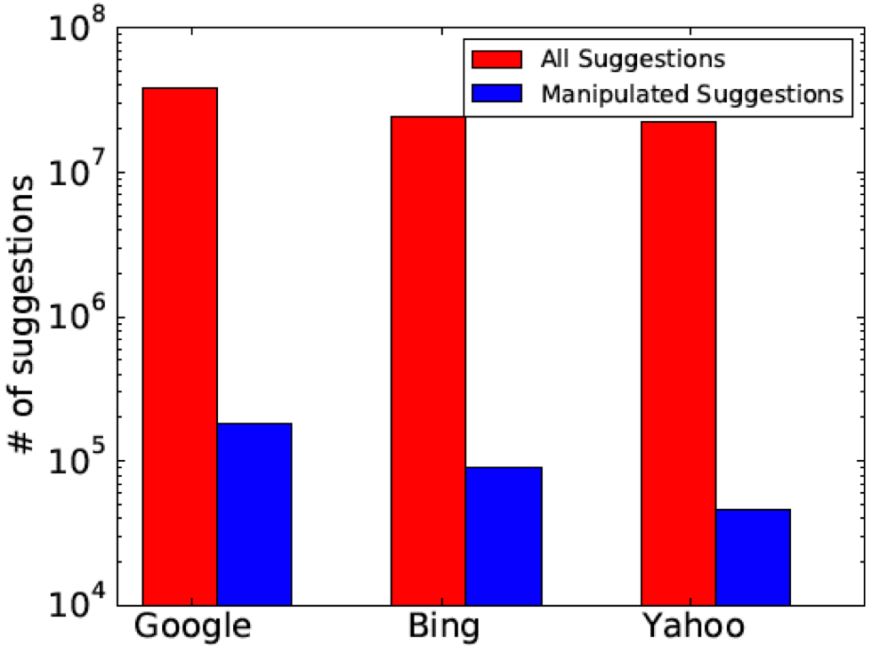
Number of missuggestions on each platform

(G: 0.48%, B: 0.37%, Y: 0.2%)

Trigger Category	# of terms	# of manipulations	Vol-ume
Lending Products	1389	1580/34629	4.13%
Home Services	17059	16712/413836	2.47%
Pharmaceutical	3715	3876/93929	2.09%
Auto Services	3477	3916/92220	1.08%
Technology	29115	32696/762548	0.91%
Education	17311	17628/423352	0.81%
Shopping	30465	32986/773087	0.72%
Gambling	413	454/10314	0.68%
Travel	6434	6554/153842	0.48%
Legal Services	14064	14084/348548	0.55%

Categories of the polluted triggers

Scope and magnitude



Number of missuggestions on each platform

(G: 0.48%, B: 0.37%, Y: 0.2%)

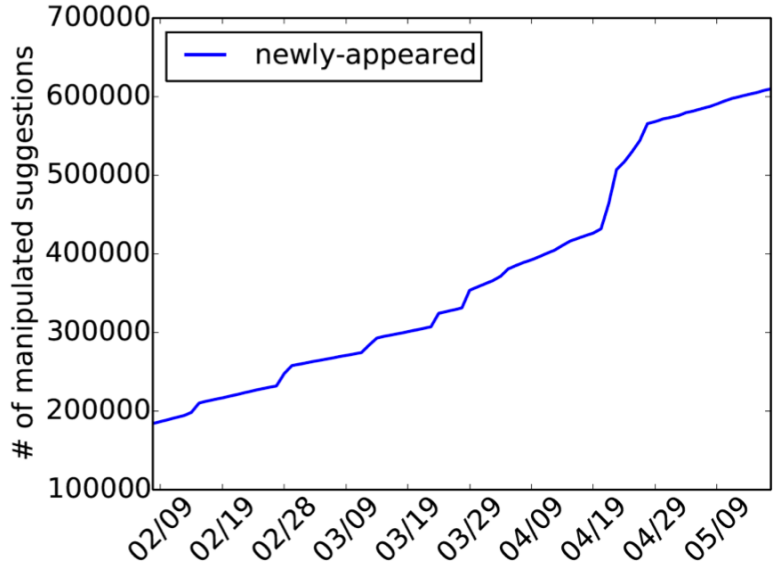
Trigger Category	# of terms	# of manipulations	Vol-ume
Lending Products	1389	1580/34629	4.13%
Home Services	17059	16712/413836	2.47%
Pharmaceutical	3715	3876/93929	2.09%
Auto Services	3477	3916/92220	1.08%
Technology	29115	32696/762548	0.91%
Education	17311	17628/423352	0.81%
Shopping	30465	32986/773087	0.72%
Gambling	413	454/10314	0.68%
Travel	6434	6554/153842	0.48%
Legal Services	14064	14084/348548	0.55%

Categories of the polluted triggers

257K polluted triggers

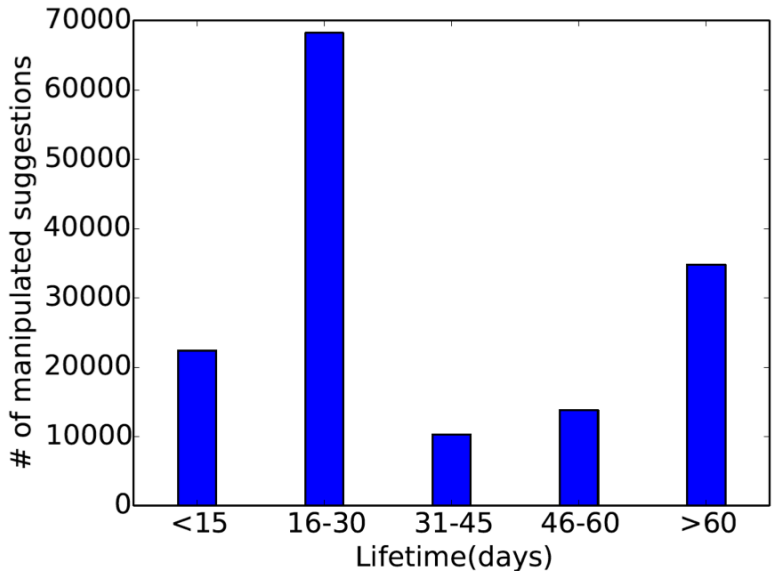
383K missuggestions

Evolution and lifetime



Number of missuggestions over time

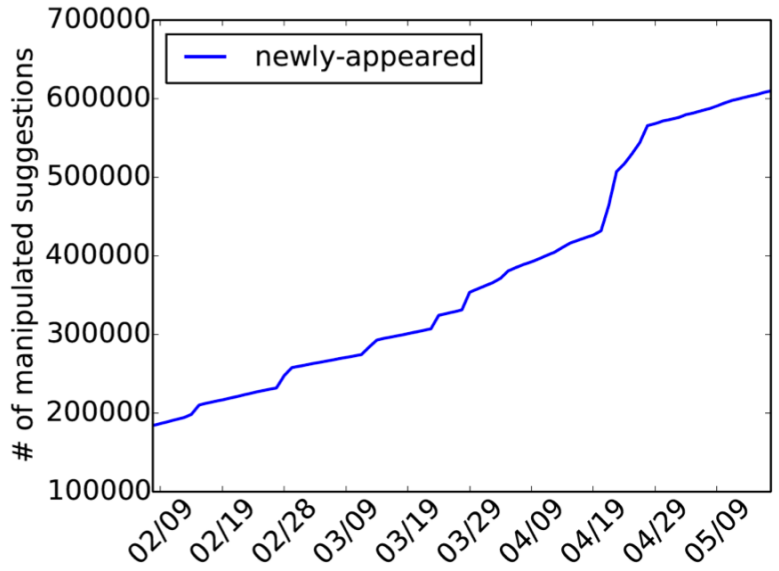
- 71.3% of newly-appeared missuggestions related to newly-appeared polluted triggers
- 1.9% of triggers were polluted on average



Lifetime distribution of missuggestions

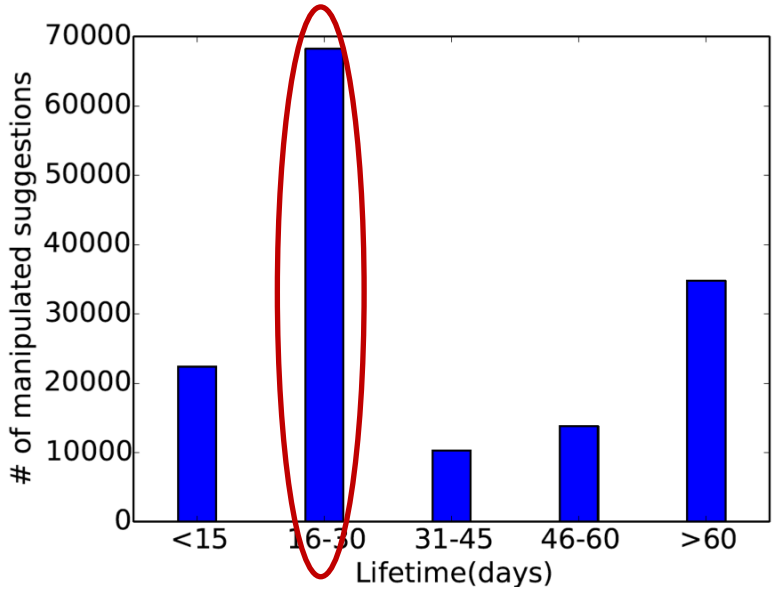
- 39.3% of missuggestions stay > 30 days
- 34 days vs. 63 days (missuggestion vs. legit.)

Evolution and lifetime



Number of missuggestions over time

- 71.3% of newly-appeared missuggestions related to newly-appeared polluted triggers
- 1.9% of triggers were polluted on average



Lifetime distribution of missuggestions

- 39.3% of missuggestions stay > 30 days
- 34 days vs. 63 days (missuggestion vs. legit.)

Missuggestion content and pattern

- 20% missuggestions related to more than one trigger

"free web hosting and domain name registration services by doteasy.com" related to 123 triggers

Missuggestion content and pattern

- 20% missuggestions related to more than one trigger

"free web hosting and domain name registration services by doteasy.com" related to 123 triggers

- missuggestion grammatical pattern

Pattern	# of missuggestions	Example
trigger relevant content+target	195,738	phoenix divorce attorney sampair
target+trigger relevant content	188,238	strongvault online backup free download
target+" - "+ trigger relevant content	2,278	bd&j - los angeles personal injury lawyers
trigger relevant content+by+URL	1,446	custom t shirt design software and application tool by panaceatek.com
trigger relevant content+from+URL	427	reliable web hosting from webhostinghub.com

Top 5 missuggestion patterns

Missuggestion content and pattern

- 20% missuggestions related to more than one trigger

"free web hosting and domain name registration services by doteasy.com" related to 123 triggers

- missuggestion grammatical pattern

Pattern	# of missuggestions	Example
trigger relevant content+target	195,738	phoenix divorce attorney sampair
target+trigger relevant content	188,238	strongvault online backup free download
target+" - "+ trigger relevant content	2,278	bd&j - los angeles personal injury lawyers
trigger relevant content+by+URL	1,446	custom t shirt design software and application tool by panaceatek.com
trigger relevant content+from+URL	427	reliable web hosting from webhostinghub.com

Top 5 missuggestion patterns

Missuggestion content and pattern

- 20% missuggestions related to more than one trigger

"free web hosting and domain name registration services by doteasy.com" related to 123 triggers

- missuggestion grammatical pattern

Pattern	# of missuggestions	Example
trigger relevant content+target	195,738	phoenix divorce attorney sampair
target+trigger relevant content	188,238	strongvault online backup free download
target+" - "+ trigger relevant content	2,278	bd&j - los angeles personal injury lawyers
trigger relevant content+by+URL	1,446	custom t shirt design software and application tool by panaceatek.com
trigger relevant content+from+URL	427	reliable web hosting from webhostinghub.com

Top 5 missuggestion patterns

Revenue analysis

- Manipulation service provider

  iXiala

  10K sites request suggestion manipulation

  $54K/week commission earned by manipulation operators

  $515K/week for 465K manipulated suggestions

Discussion

- Limitations
  - adversary can make the manipulations mimic benign ones
  - lack of ground truth, manual efforts involved

Discussion

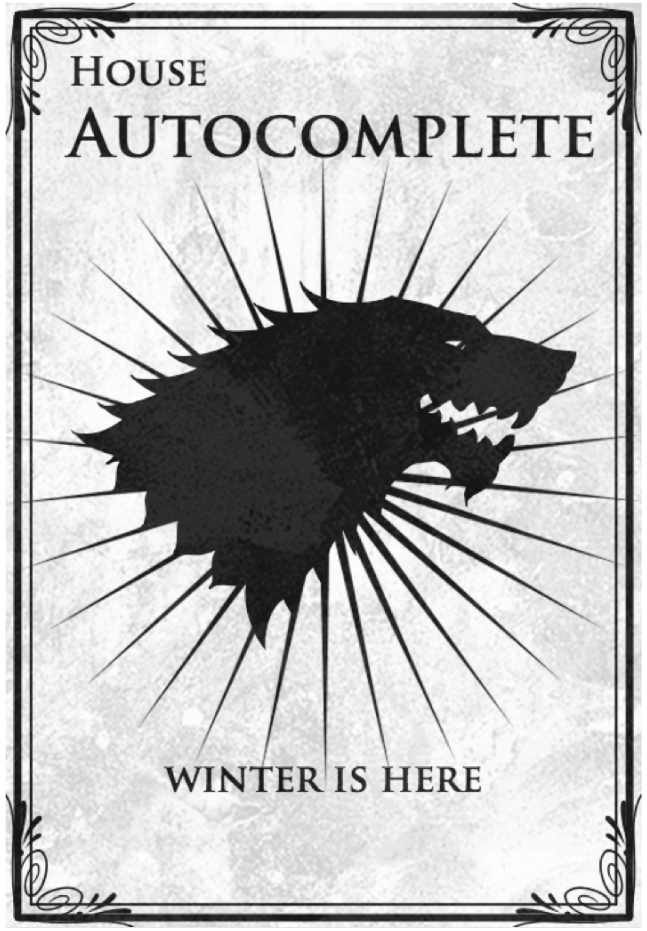
- Limitations
  - adversary can make the manipulations mimic benign ones
  - lack of ground truth, manual efforts involved
- Lesson learned
  - unpopular targets related to triggers
  - similar keyword patterns

Conclusion

- **first large-scale analysis** of autocomplete missuggestions, and make **first step** to understand the underground ecosystem

- **novel NLP techniques** to build up the first detection system **without accessing** to search logs

Questions & Answers



Data collection

- Datasets

Dataset	# of suggestions	# of triggers	# of result pages
Badset	150	145	295
Goodset	300	298	593
Unknown set	114,275,000	1,000,900	1,607,951

- Validation criteria
  - missuggestion must promote a target whose own reputation cannot make itself stand out in the search results of the trigger
  - missuggestion and its search results conflict with the user's original search intention

Semantic Consistency Classifier

- 100 missuggestions + 150 legitimate trigger-suggestion pairs
- SVM classification model with 5-folder cross validation
- Precision 94.59%, Recall 95.89%

Label	Feature	F-score
$F_{ss}(s^a, s^t)$	sentence similarity	0.597
$F_{ws}(w^a, w^t)$	word similarity	0.741
$F_{if}(w^a, w^t)$	infrequency	0.653

Missuggestion Classifier

- 150 missuggestions + 300 legitimate trigger-suggestion pairs
- SVM classification model with 5-folder cross validation
- Precision: 96.23%, Recall 95.63%

Label	Feature	F-score
$F_{rs}(D^a, D^t)$	result similarity	0.782
$F_{ci}(w^a, H^a, H^t)$	content impact	0.808
$F_{\mathrm{rp}}(D^a, D^t)$	result popularity	0.632
$F_{rs}(N^a, N^t)$	result size	0.745

Evaluation

- Accuracy and coverage
  - Tow-step analysis : precision 96.23%, recall 95.63%on ground truth
  - One-step analysis: precision 97.68%, recall 95.59% on ground truth

- Performance
  - Tow-step analysis: 0.016s/pair (94X faster)
  - One-step analysis: 1.5s/pair